Penalized nonparametric scalar-on-function regression via principal coordinates

Philip T. Reiss*
Department of Child and Adolescent Psychiatry
and Department of Population Health, New York University,
and Department of Statistics, University of Haifa

David L. Miller
Centre for Research into Ecological and Environmental Modelling
and School of Mathematics and Statistics,
University of St Andrews, St Andrews, Scotland, United Kingdom

Pei-Shien Wu and Wen-Yu Hua
Department of Child and Adolescent Psychiatry, New York University

November 24, 2015

Abstract

A number of classical approaches to nonparametric regression have recently been extended to the case of functional predictors. This paper introduces a new method of this type, which extends intermediate-rank penalized smoothing to scalar-on-function regression. The core idea is to regress the response on leading principal coordinates defined by a relevant distance among the functional predictors, while applying a ridge penalty. Our publicly available implementation, based on generalized additive modeling software, allows for fast optimal tuning parameter selection and for extensions to multiple functional predictors, exponential family-valued responses, and mixed-effects models. In an application to signature verification data, the proposed principal coordinate ridge regression is shown to outperform a functional generalized linear model.

Keywords: dynamic time warping, functional regression, generalized additive model, kernel ridge regression, multidimensional scaling

*Philip Reiss, Pei-Shien Wu and Wen-Yu Hua gratefully acknowledge the support of the U.S. National Institute of Mental Health (grant 1R01MH095836-01A1).

1 Introduction

A central problem in functional data analysis is to relate scalar responses y_i to functional predictors $x_i(\cdot)$ $(i = 1, \ldots, n)$ by a regression model. In the terminology of Reiss et al. (2010), models of this type are known as "scalar-on-function" regression, to distinguish them from models for functional responses ("function-on-scalar" or "function-on-function" regression). The standard approach (Marx and Eilers, 1999; Cardot et al., 1999; Ramsay and Silverman, 2005) is to estimate the intercept α and slope or coefficient function $\beta(\cdot)$ in the functional linear model

$$y_i = \alpha + \int_{\mathcal{T}} x_i(t)\beta(t)dt + \varepsilon_i \tag{1}$$

where $\mathcal{T}$ is the domain of the functional predictors and $E(\varepsilon_i) = 0$.

But as with ordinary (scalar-predictor) regression, a linear model can sometimes fail to capture the relationship of interest. As an illustration, suppose that $x_i : [0, 1] \longrightarrow \mathbb{R}$, the ith of n functional predictors, is a noisy realization of

$$x_i^0(t) = \left\{ \begin{array}{ll} -1, & t - \tau_i \in [-.05, 0); \\ 1, & t - \tau_i \in [0, .05); \\ 0, & \text{otherwise}, \end{array} \right\} \text{ for some } \tau_i \in [0, 1], \tag{2}$$

as in the first two panels of Figure 1. Suppose further that we have responses arising as

$$y_i = \tau_i + \varepsilon_i \tag{3}$$

with independent identically distributed mean-zero normal errors ε_i. In this setup, y_i depends strongly on $x_i(\cdot)$ via the parameter τ_i which characterizes the latter. This dependence is not captured by the linear model (1).

On the other hand, if we can define a distance between functional predictors such that $x_i(\cdot), x_j(\cdot)$ are close when τ_i, τ_j are, then intuitively, the principle that similar $x(\cdot)$'s imply similar y's may be a better guide to extracting the information in $x(\cdot)$ relevant to predicting y. Thus a predictive algorithm based on this principle may succeed where the functional linear model fails.

Motivated by this idea of exploiting a relevant distance among the functional predictors to predict the response, this paper proposes a simple but powerful new approach, which we

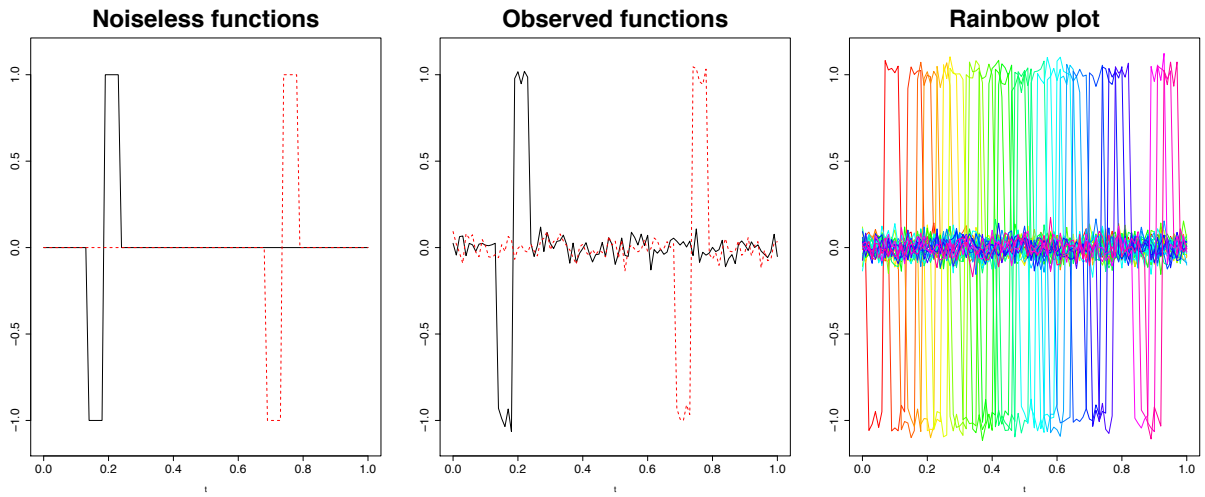

Figure 1: Two instances of the noiseless functional predictors x_i^0 (2) in the toy example are shown in the left panel, with corresponding observed functions x_i shown in the middle panel. The right panel shows $n = 30$ such functional predictors, color-coded as in Hyndman and Shang (2010) by the corresponding responses (3).

call principal coordinate ridge regression (PCoRR), for scalar-on-function regression. As a brief illustration of PCoRR's utility, we used windowed dynamic time warping (DTW; Sakoe and Chiba, 1978; Giorgino, 2009; Faraway, 2012) to define distances among the functional predictors shown in Figure 1. For this data set, the windowed DTW distance (see §4) captures the information relevant to prediction of y. Consequently, as shown in Figure 2, the proposed method, using principal coordinates based on that distance, dramatically outperforms an analogous functional linear model.

A similar motivation underlies previous methodology for nonparametric scalar-on-function regression (e.g., Ferraty and Vieu, 2006). But unlike previous proposals, our approach can be implemented using existing software for generalized additive and related models (Wood, 2006, 2011). This allows PCoRR to be extended readily to the wide range of models available with such software, including models with generalized linear responses, multiple functional predictors, and/or random effects.

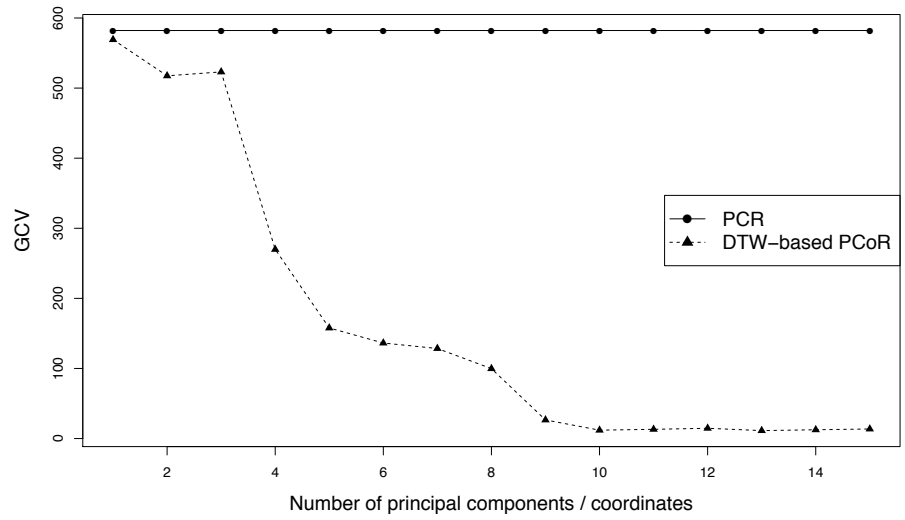

Figure 2: Performance for the toy data: ridge regression on principal component scores, a linear approach to scalar-on-function regression, vs. on (DTW-based) principal coordinates, the proposed nonparametric approach. The latter attains much lower (better) generalized cross-validation scores (see §2.5).

2 Methodology

2.1 Principal coordinates

Let $\boldsymbol{D} = (d_{ij})_{1 \le i,j \le n}$ be a symmetric matrix of distances with $d_{ii} = 0$ for all i and $d_{ij} \ge 0$ for all i, j. Classical multidimensional scaling (Gower, 1966) seeks n points in $\mathbb{R}^q$ (for some $q \le n$) whose Euclidean distances are "closest" to $\boldsymbol{D}$, in the sense discussed in section 14.4 of Mardia et al. (1979).

Define the $n \times n$ centering matrix $\boldsymbol{H} = \boldsymbol{I}_n - \mathbf{1}_n \mathbf{1}_n^T / n$ and let

$$\boldsymbol{K} = \boldsymbol{H} \left(-\frac{1}{2} d_{ij}^2 \right)_{1 \le i,j \le n} \boldsymbol{H} \tag{4}$$

have leading eigenvalues $\lambda_1 \ge \ldots \ge \lambda_q > 0$, with corresponding eigenvectors $\boldsymbol{w}_{\cdot 1}, \ldots, \boldsymbol{w}_{\cdot q}$ scaled so that $\|\boldsymbol{w}_{\cdot \ell}\|^2 = \lambda_\ell$ for $\ell = 1, \ldots, q$. Then the desired n points are given by the *rows* of the $n \times q$ matrix $\boldsymbol{W}_q = (\boldsymbol{w}_{\cdot 1} \ \ldots \ \boldsymbol{w}_{\cdot q})$. These row vectors, which we may denote by $\boldsymbol{w}_{1\cdot}, \ldots, \boldsymbol{w}_{n\cdot}$, are called the *principal coordinates* (PCo's) of the data.

To construct the PCo matrix $\boldsymbol{W}_q$ explicitly, consider the eigendecomposition

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}^T \tag{5}$$

where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_n$ and $\boldsymbol{\Delta} = \text{Diag}\{\delta_1, \ldots, \delta_n\}$ with $\delta_1 \geq \ldots \geq \delta_n$. Let $\boldsymbol{U} = (\boldsymbol{U}_q\ \boldsymbol{U}_{-q})$ and $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\Delta}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Delta}_{-q} \end{pmatrix}$ where $\boldsymbol{U}_q$ is $n \times q$ and $\boldsymbol{\Delta}_q$ is $q \times q$, and assume $\delta_q > 0$. From the above definition, we have

$$\boldsymbol{W}_q = \boldsymbol{U}_q\boldsymbol{\Delta}_q^{1/2}. \tag{6}$$

2.2 Ridge regression on principal coordinates

Given a distance defined among the n instances of the functional predictor, our proposal is simply to perform ridge regression on the leading q PCo's with respect to this distance, for a chosen $q > 0$. More explicitly, suppose we have a response vector $\boldsymbol{y} \in \mathbb{R}^n$, an $n \times p$ design matrix $\boldsymbol{M}$ of scalar covariates, and the $n \times q$ matrix $\boldsymbol{W}_q$ of PCo's, with $p + q < n$. The basic proposal for PCoRR is the model fit

$$\hat{\boldsymbol{y}} = \boldsymbol{M}\hat{\boldsymbol{\alpha}} + \boldsymbol{W}_q\hat{\boldsymbol{\gamma}} \tag{7}$$

where $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ minimizes the ridge-type criterion

$$\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha} - \boldsymbol{W}_q\boldsymbol{\gamma}\|^2 + \lambda\boldsymbol{\gamma}^T\boldsymbol{\gamma} \tag{8}$$

for some $\lambda > 0$.

When $\boldsymbol{M}$ is of full rank and orthogonal to the principal coordinates ($\boldsymbol{M}^T\boldsymbol{W}_q = \boldsymbol{0}$), the fitted value matrix has the explicit form

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{M}(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{y} + \boldsymbol{U}_q\text{Diag}\left\{\frac{\delta_\ell}{\delta_\ell + \lambda}\right\}_{1\leq\ell\leq q}\boldsymbol{U}_q^T\boldsymbol{y} \\ &= \boldsymbol{M}(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{y} + \sum_{1\leq\ell\leq q}\frac{\delta_\ell}{\delta_\ell + \lambda}\boldsymbol{u}_\ell\boldsymbol{u}_\ell^T\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{u}_\ell$ is the ℓth column of $\boldsymbol{U}$. In other words, the projection of $\boldsymbol{y}$ on the ℓth PCo direction is multiplied by the "filter factor" $\frac{\delta_\ell}{\delta_\ell+\lambda}$ (Randolph et al., 2012), with preference (larger factors, or less shrinkage) given to directions corresponding to larger eigenvalues, resulting in a fit with effective degrees of freedom $p + \sum_{\ell=1}^q \frac{\delta_\ell}{\delta_\ell+\lambda}$.

2.3 Prediction for new observations

Expression (7) tells us only the fitted values for the n given observations ("design points"). Suppose we have n^* new observations of the functional predictor, along with an $n^* \times p$ matrix $\boldsymbol{M}^*$ of covariate values. To obtain predictions for these non-design points we must first "insert" them into the PCo configuration. In other words, we need an $n^* \times q$ matrix $\boldsymbol{W}_q^*$ of coordinates with respect to this configuration, which would yield predicted values $\hat{\boldsymbol{y}}^* = \boldsymbol{M}^*\hat{\boldsymbol{\alpha}} + \boldsymbol{W}_q^*\hat{\boldsymbol{\gamma}}$. A solution is given by applying Gower's interpolation (Gower, 1968) to each point: letting d_{mi}^* denote the distance from the mth new point to the ith design point, and k_{ii} the ith diagonal element of $\boldsymbol{K}$ in (4), we take

$$\boldsymbol{W}_q^* = \boldsymbol{K}^*\boldsymbol{U}_q\boldsymbol{\Delta}_q^{-1/2}, \tag{9}$$

where $\boldsymbol{K}^*$ is the $n^* \times n$ matrix with (m, i) entry $-\frac{1}{2}(d_{mi}^{*2} - k_{ii})$. Subtracting k_{ii} is a form of centering (Gower, 1968), and we use the notation $\boldsymbol{K}^*$ to suggest that this matrix is constructed by transforming the distances d_{mi}^* in a roughly similar manner to the transformation of the original distance matrix $\boldsymbol{D}$ to obtain $\boldsymbol{K}$. See Appendix A for further discussion.

2.4 Extending the model

Generalized additive modeling software, specifically the `mgcv` package (Wood, 2006, 2011) for R (R Core Team, 2015), allows for several important extensions to the PCoRR fit (7):

1. A generalized linear model (GLM) extension is achieved by minimizing the penalized deviance $D(\boldsymbol{\alpha}, \boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}^T\boldsymbol{\gamma}$, as opposed to the penalized sum of squares (8).

2. Since `mgcv` can perform fast optimal selection of multiple smoothing parameters, one can incorporate multiple functional predictors. Letting $\boldsymbol{W}_{q_r}^{(r)}$ denote the PCo matrix for the rth of R functional predictors, criterion (8) is then replaced by

$$\left\| \boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha} - \sum_{r=1}^{R} \boldsymbol{W}_{q_r}^{(r)}\boldsymbol{\gamma}_r \right\|^2 + \sum_{r=1}^{R} \lambda_r \boldsymbol{\gamma}_r^T\boldsymbol{\gamma}_r \tag{10}$$

   (with sum of squared errors again replaced by deviance for the generalized linear case).

3. Ruppert et al. (2003), Reiss and Ogden (2009) and Wood (2011), among others, propose to choose the penalty parameter λ (or parameters $\lambda_1, \ldots, \lambda_R$) by maximum likelihood or restricted maximum likelihood (REML), or approximate versions thereof in the generalized linear case. Since these criteria are also the standard procedures for fitting mixed-effect models, their use allows for straightforward incorporation of random effects in the model.

2.5 Tuning parameter selection

In (10), we must choose both the dimension q_r and the ridge penalty parameter λ_r, for $r = 1, \ldots, R$. As noted above, we favor maximum likelihood or REML for the choice of λ_r. For q_r, in line with Wood (2006), we use generalized cross-validation (GCV; Craven and Wahba, 1979) for GLMs with unknown scale parameter (e.g., linear regression), and the Akaike information criterion (AIC; Akaike, 1973) for those with known scale parameter (e.g., logistic regression). Typically there is a minimum number of coordinates above which these criteria attain an approximate plateau (Miller and Wood, 2014), much as has been found regarding the number of knots for penalized splines (Ruppert, 2002). Consequently, when $R > 1$ it seems adequate to fix $q_1 = \ldots = q_R = q$ for a sufficiently high value of q, perhaps 10 or 20, rather than painstakingly optimizing GCV or AIC over $(q_1, \ldots, q_R)$.

2.6 Implementation

We have implemented PCoRR in the R package `poridge` (Principal co-Ordinate Ridge regression), which is publicly available at `https://github.com/dill/poridge`. Essentially the package is an add-on to the `mgcv` package (Wood, 2006, 2011). It implements a non-standard "smooth constructor" function that inputs a distance matrix $\boldsymbol{D}$ and a number of coordinates q, and enables the workhorse function `s()` to add the resulting term $+\boldsymbol{W}_q\boldsymbol{\gamma}$ to a generalized additive model, along with a ridge penalty.

3 Relationships with other work

A number of previous approaches in the functional data literature and elsewhere have connections with PCoRR, some of which may not be immediately apparent. This section explains some of these relationships. Situating our method in relation to others may offer new insights and may help to suggest alternative avenues for future work.

3.1 Functional principal component regression

To relate our method to functional principal component regression, we first need to recall the notion of PCo-PCA "duality" developed by Gower (1966) for multivariate data. Given an $n \times p$ matrix $\boldsymbol{X}$ of n observations in $\mathbb{R}^p$, the key result is that the principal coordinates arising from Euclidean distances among the observations are equal to the principal component scores. We wish to generalize this to observations $x_1, \ldots, x_n$ in a separable Hilbert space $\mathcal{F}$, such as $L^2(\mathcal{T})$ or a subspace thereof, equipped with an inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle}$.

The proof of duality for multivariate data relies on the one-to-one correspondence between eigenvectors of $\boldsymbol{X}_c \boldsymbol{X}_c^T$ and those of $\boldsymbol{X}_c^T \boldsymbol{X}_c$, where $\boldsymbol{X}_c = \boldsymbol{H}\boldsymbol{X}$. To generalize the result, we need to introduce an abstract counterpart of the matrix $\boldsymbol{X}_c$, namely the bounded linear transformation $T_x : \mathcal{F} \longrightarrow \mathbb{R}^n$ defined by $T_x f = (\langle x_1 - \bar{x}, f \rangle, \ldots, \langle x_n - \bar{x}, f \rangle)^T$. It is readily checked that the adjoint of T_x is $T_x^* : \mathbb{R}^n \longrightarrow \mathcal{F}$ given by $T_x^* \boldsymbol{v} = \sum_{i=1}^n (v_i - \bar{v}) x_i$ where $\boldsymbol{v} = (v_1, \ldots, v_n)^T$. We then have, for $f \in \mathcal{F}$,

$$T_x^* T_x f = \sum_{i=1}^n \langle x_i - \bar{x}, f \rangle (x_i - \bar{x}) = \sum_{i=1}^n \langle x_i - \bar{x}, f \rangle x_i,$$

i.e., $T_x^* T_x$ is $n-1$ times the sample covariance operator on $\mathcal{F}$; and for $\boldsymbol{v} \in \mathbb{R}^n$, $T_x T_x^* \boldsymbol{v} = \boldsymbol{H}\boldsymbol{C}\boldsymbol{H}\boldsymbol{v}$ where $\boldsymbol{C} = (\langle x_i, x_j \rangle)_{1 \le i,j \le n}$.

$T_x^* T_x$ has eigenexpansion

$$T_x^* T_x f = \sum_{\ell=1}^{n-1} \delta_\ell \langle \phi_\ell, f \rangle \phi_\ell \tag{11}$$

where $\delta_1 \ge \delta_2 \ge \ldots \ge 0$ and the ϕ_ℓ's are orthonormal. One way to see that $n-1$ terms suffice in this expansion is to observe that for each eigenvalue-eigenelement pair (δ_ℓ, ϕ_ℓ) of

$T_x^* T_x$, $\boldsymbol{HCH}$ has eigenvalue δ_ℓ since

$$\boldsymbol{HCH}(T_x\phi_\ell) = (T_x T_x^*)T_x\phi_\ell = T_x(T_x^* T_x)\phi_\ell = \delta_\ell T_x\phi_\ell; \tag{12}$$

but $\boldsymbol{HCH}\mathbf{1}_n = \mathbf{0}$, so $\boldsymbol{HCH}$ has at most $n-1$ positive eigenvalues. The orthonormal eigenelements $\phi_1, \phi_2, \ldots, \phi_{n-1}$ form what we may call a (sample) $\mathcal{F}$-principal component basis of $\mathcal{F}$. The ℓth $\mathcal{F}$-PC scores for the data points are given by $\langle x_i - \bar{x}, \phi_\ell \rangle : i = 1, \ldots, n$, i.e., by $T_x\phi_\ell$. For $\mathcal{F}$ a space of square-integrable functions, the $\mathcal{F}$-PC expansion is essentially the FPC expansion as formulated by, for example, Dauxois et al. (1982) and Aguilera et al. (1999). We can now state our duality theorem for Hilbert space-valued data.

Theorem 1. *Assume a set of observations $x_1, \ldots, x_n$ in a separable Hilbert space $\mathcal{F}$ satisfies $\delta_1 > \ldots > \delta_q > 0$ in (11) where $q \in \{1, \ldots, n-1\}$. Then for $\ell = 1, \ldots, q$, the vector of ℓth principal coordinates with respect to the distance matrix $\boldsymbol{D} = (\|x_i - x_j\|_{\mathcal{F}})_{1 \le i,j \le n}$ is equal, with a possible sign change, to the vector of scores with respect to the ℓth $\mathcal{F}$-PC.*

Proof. Letting $c_{ij} = \langle x_i, x_j \rangle$ and $d_{ij} = \|x_i - x_j\|_{\mathcal{F}}$, we have

$$d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}, \tag{13}$$

i.e., $\boldsymbol{D}$ is obtained from $\boldsymbol{C}$ by the "standard transformation" of Mardia et al. (1979), p. 402. Hence by Theorem 14.2.2 of Mardia et al. (1979), $\boldsymbol{HCH}$ equals the matrix $\boldsymbol{K}$ of (4). By the distinctness of the leading q eigenvalues, it suffices to show (i) that $T_x\phi_\ell$, the vector of scores with respect to the ℓth $\mathcal{F}$-PC, is an eigenvector of $\boldsymbol{HCH}$ corresponding to that matrix's ℓth largest eigenvalue δ_ℓ; and (ii) that $T_x\phi_\ell$ has squared norm δ_ℓ. But we have already established (i) in (12), and (ii) is clear since $\|T_x\phi_\ell\|^2 = \phi_\ell^T(T_x^* T_x\phi_\ell) = \delta_\ell(\phi_\ell^T\phi_\ell) = \delta_\ell$. □

If $\mathcal{F}$ is a space of square-integrable functions on $\mathcal{T}$, says that principal coordinates with respect to the usual L^2 distance $\|f - g\| = \sqrt{\int_{\mathcal{T}}[f(t) - g(t)]^2 dt}$ are simply FPC scores. Thus regressing on PCo's with respect to an *arbitrary* distance among functions is a direct generalization of functional principal component regression.

In light of Theorem 1, we can view several FPC-based methods for scalar-on-function regression, referred to in the first row of Table 1, as being generalized by the corresponding PCo-based methods in the second row. The first column of Table 1 refers to unpenalized

Basis type	Linear (unpenalized)	Ridge (penalized)	Additive
FPC	Cardot et al. (1999)	Reiss and Ogden (2007)	Müller and Yao (2008)
PCo	Cuadras et al. (1996)	PCoRR	—

Table 1: Selected references for some approaches to scalar-on-function regression via functional principal component bases (first row) and principal coordinate bases (second row).

linear regression on leading FPC scores, a popular approach to scalar-on-function regression (e.g., Cardot et al., 1999; Müller and Stadtmüller, 2005), as a special case of regression on leading PCo's. The latter method has been discussed extensively by Cuadras and Arenas (1990) and Cuadras et al. (1996), who refer to it as "distance-based regression," and has been recently applied to functional predictors (Boj et al., 2015). In contrast to these unpenalized methods, the methods in the second column of Table 1 incorporate a penalty: the "FPCR$_R$" method of Reiss and Ogden (2007) regresses on leading FPCs with a roughness penalty, whereas PCoRR adopts a ridge penalty (and thus is not an exact generalization of FPCR$_R$). As in the case of scatterplot smoothing, introducing a penalty enables one to use a richer basis, and the simulation results of Reiss and Ogden (2007) suggest that this yields improved performance for scalar-on-function regression (see the discussion of Horváth and Kokoszka, 2012). To be sure, Cuadras et al. (1996) also consider extensions of simple linear regression, including ridge regression; but our PCoRR formulation, which achieves the extensions outlined in §2.4 with automatically optimized tuning parameters, is new. The third column of Table 1 refers to additive models. These have been developed for FPC scores by Müller and Yao (2008), and might be studied for general PCo's in future work.

3.2 Nonparametric functional regression

A more flexible alternative to the functional linear model (1) is the nonparametric model

$$y = m(x) + \varepsilon, \tag{14}$$

where m is some mapping from the function space of interest to $\mathbb{R}$ and $E(\varepsilon|x) = 0$. To estimate m, one may extend nonparametric regression methodology from the case of scalar

If x is...	Kernel smoothing	Local polynomial	Full-rank penalized	Reduced-rank penalized
Scalar	Nadaraya (1964); Watson (1964)	Fan and Gijbels (1996)	Wahba (1990)	Eilers and Marx (1996); Ruppert et al. (2003); Wood (2003)
Functional	Ferraty and Vieu (2006)	Baíllo and Grané (2009)	Preda (2007)	PCoRR

Table 2: Selected references for some approaches to nonparametric regression with scalar predictors (first row) and functional predictors (second row).

x to the case of functional x (Geenens, 2011). For example, the original proposal for nonparametric scalar-on-function regression (Ferraty and Vieu, 2006) was a functional version of the kernel smoothing method of Nadaraya (1964) and Watson (1964). Likewise, Baíllo and Grané (2009) extended local linear smoothing to the functional predictor setting.

Somewhat less transparently, roughness penalty approaches to nonparametric regression also have functional-predictor counterparts, as suggested by the two rightmost columns of Table 2. Given observations $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$, the smoothing spline approach to nonparametric regression (Wahba, 1990) posits a model of the form (14), and finds the estimator $\hat{m}$ that minimizes a penalized loss over a reproducing kernel Hilbert space $\mathcal{H}$ of functions $m : \mathbb{R} \longrightarrow \mathbb{R}$. Considering for simplicity the case of squared error loss and $\|\cdot\|_{\mathcal{H}}$ having trivial null space, the smoothing spline estimator is

$$\hat{m} = \operatorname*{arg\,min}_{m \in \mathcal{H}} \left[\sum_{i=1}^{n} \{y_i - m(x_i)\}^2 + \lambda \|m\|_{\mathcal{H}}^2 \right] \tag{15}$$

for some $\lambda > 0$. The nonparametric functional regression estimator of Preda (2007) has the same form, but here $x_1, \ldots, x_n$ belong to a function space $\mathcal{F}$, and $\mathcal{H}$ is an RKHS of maps $m : \mathcal{F} \longrightarrow \mathbb{R}$. Briefly, the RKHS $\mathcal{H}$ considered by Preda (2007) is generated by a kernel $k : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$ such as $k(f, g) = \exp(-\|f - g\|_{\mathcal{F}}^2 / 2\sigma^2)$ ($\sigma > 0$), in the sense that $\mathcal{H}$ is the completion of

$$\left\{ \sum_{\ell=1}^{L} a_\ell k(\cdot, f_\ell) : a_1, \ldots, a_L \in \mathbb{R}, f_1, \ldots, f_L \in \mathcal{F} \right\}$$

with respect to a specific inner product. By the representer theorem (Schölkopf et al., 2001), the minimizer in (15) is of the form $m(x) = \sum_{i=1}^{n} \beta_i k(x, x_i)$. For $\hat{m}$ of this form the

right side of (15) can be shown to equal

$$\|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{K} \boldsymbol{\beta} \tag{16}$$

where

$$\boldsymbol{K} = [k(x_i, x_j)]_{1 \leq i,j \leq n} \tag{17}$$

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$. Thus $\hat{m}(x) = \sum_{i=1}^n \hat{\beta}_i k(x, x_i)$ where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)^T$ minimizes (16). In the machine learning literature this is referred to as kernel ridge regression (KRR; e.g., Shawe-Taylor and Cristianini, 2004).

To place PCoRR in the context of nonparametric functional regression we must proceed to the fourth column of Table 2. In the top entry of that column we find penalized intermediate-rank splines (e.g., Green and Silverman, 1994; Ruppert et al., 2003; Wood, 2006). In comparison to smoothing splines, these approaches can be viewed as approximate RKHS optimizers offering greater computational efficiency. Similarly PCoRR can be thought of as approximating the RKHS approach of Preda (2007) to functional regression. To see this, suppose that $\boldsymbol{K}$ in (16) arises not as the Gram matrix (17), but from a distance matrix as in (4). (See Appendix B for discussion of the correspondence between distance and kernel matrices.) Suppose we impose the the restriction

$$\boldsymbol{\beta} = \boldsymbol{U}_q \boldsymbol{\Delta}_q^{-1/2} \boldsymbol{\gamma} \tag{18}$$

for some $\boldsymbol{\gamma} \in \mathbb{R}^q$. Substituting into (16), and using (5) and (6), yields the criterion $\|\boldsymbol{y} - \boldsymbol{W}_q \boldsymbol{\gamma}\|^2 + \lambda \boldsymbol{\gamma}^T \boldsymbol{\gamma}$, which is simply the PCoRR criterion (8) (where the covariate part $\boldsymbol{M}\boldsymbol{\alpha}$ has been omitted for simplicity). Since restriction (18) confines $\boldsymbol{K}\boldsymbol{\beta}$ to the column space of $\boldsymbol{U}_q$, an optimal q-dimensional approximation in the sense of Eckart and Young (1936) (as in Reiss and Ogden, 2010), PCoRR can be viewed as optimally truncated KRR. PCoRR, then, seems particularly reminiscent of thin plate regression splines (Wood, 2003), which are essentially thin plate splines with a similar optimal truncation step. At any rate, our aim here, in highlighting the analogy between PCoRR and intermediate-rank spline smoothing, has been to present our method as a natural, and potentially fruitful, next step in the development of nonparametric scalar-on-function regression.

3.3 Other kernel methods

Our interpretation of PCoRR as truncated KRR implies that multiple-term models, such as the optimizer of (10), constitute a form of *multiple kernel learning* (see Gönen and Alpaydın, 2011, for a review). The standard viewpoint of multiple kernel learning differs from ours in several respects, however: one starts with kernels $k_1(\cdot,\cdot),\ldots,k_R(\cdot,\cdot)$ as opposed to distances, the aim is usually classification (as opposed to general exponential family outcomes), and the idea is to combine the R kernels into a single kernel in some optimal manner.

The viewpoint of Corrada Bravo et al. (2009) is in a sense intermediate between ours and that of the kernel literature. Like us, these authors begin not with a kernel, but with a distance function that may not be Euclidean and need not even satisfy the triangle inequality (although they reserve the term "distance" for functions that do satisfy the triangle inequality, and use "dissimilarity" for those that do not). But as in most of the kernel literature, Corrada Bravo et al. (2009) require a positive semidefinite matrix $\boldsymbol{K}$. Thus the "pseudo-attributes" entered into their smoothing spline ANOVA model have the same form as the PCo's (6), but arise from a positive semidefinite $\boldsymbol{K}$ obtained by convex cone programming (Lu et al., 2005) rather than from (4). A less computationally intensive way to derive a positive semidefinite $\boldsymbol{K}$ from a non-Euclidean distance matrix $\boldsymbol{D}$ is to add the constant derived by Cailliez (1983) to the non-diagonal entries of $\boldsymbol{D}$, and then apply (4); this approach is available as an option in our PCoRR implementation.

4 Application: Signature verification

4.1 The data set

We now consider part of the sample data from the First International Signature Verification Competition (SVC 2004), available at `http://www.cse.ust.hk/svc2004/`. Each individual in the sample contributed 20 genuine signatures, which were accompanied by 20 skilled forgeries. For each signature we have x- and y-coordinates recorded at $\approx$ 150–300 time points. The challenge is to design an algorithm that can distinguish the genuine signatures from the fakes. Geenens (2011) considers a functional NW estimator for these

Figure 3: Five instances of a Chinese signature (above) along with five skilled forgeries thereof (below).

data. For illustration we consider the data for one individual. Figure 3 displays five of the 20 instances of this individual's Chinese signature, along with five of the corresponding forgeries.

4.2 Dynamic time warping

Noting that dynamic time warping approaches are considered state-of-the-art for signature verification (Kholmatov and Yanikoglu, 2005; Houmani et al., 2012), we apply DTW-based PCoRR to this data set. Briefly, DTW refers to a set of dynamic programming algorithms for optimally aligning discretely observed functions $[f(s), s = 1, \ldots, S$ and $g(t), t = 1, \ldots, T$, by choosing pairs $(s_1, t_1), \ldots, (s_K, t_K)$ to minimize a normalized weighted average of

$$|f(s_k) - g(t_k)|, \tag{19}$$

subject to a set of constraints including (i) $s_1 = t_1 = 1$, $s_K = S$, $t_K = T$ and (ii) $s_{k+1} - s_k, t_{k+1} - t_k \in \{0, 1\}$ for each k. That weighted average, referred to as the DTW distance, need not be a metric. Another possible constraint is $|s_k - t_k| \leq W$ for some $W > 0$ (Sakoe and Chiba, 1978); this was used in the toy example of the introduction to ensure that pairs of curves with nearby values of τ would have low distance.

4.3 Comparison

For this illustration we compare logistic PCoRR based on the DTW distance versus logistic functional PCR (a functional GLM) with a ridge penalty. We implemented FPCR as PCoRR based on L^2 distance (see §3.1) to eliminate irrelevant differences in implementation between FPCR and DTW-based PCoRR. Both approaches perform "online" signature verification, meaning that they use pen timing information in addition to the shape of the signature.

To implement FPCR, we transformed each signature's time scale to the interval $[0, 1]$. Using the R package `fda` (Ramsay et al., 2014), we set up a basis of 30 quintic B-spline functions on $[0, 1]$, with which we performed penalized smoothing of each signature's x- and y-coordinates, and also obtained estimates of the first and second derivatives of the x- and y-curves. We then formed distance matrices between 1) the original curves, 2) the first derivatives and 3) the second derivatives; all three distance matrices were attempted as inputs to PCoRR. The smoothed x- and y-curves and their derivative estimates are displayed in Figure 4. One of the fake signatures is an outlier, as can be seen clearly both in the second row of the figure and in the scatterplots of first vs. second FPC scores in the third row.

For nonparametric (DTW-based, as opposed to L^2-based) PCoRR, we calculated DTW distances among the original signatures and first- and second-differenced versions thereof, using the R package `dtw` (Giorgino, 2009). For bivariate time series such as these, Euclidean distance in $\mathbb{R}^2$ is substituted for absolute value in (19) to define the DTW distance. First vs. second PCo's are plotted for the original data and for the differenced data in Figure 5. The scatterplots suggest that the leading DTW-based PCo's for both the first- and the second-differenced data do a good job of separating the two groups. Consistent with this, the lower right subfigure shows that logistic PCoRR based on DTW distances for the differenced data attains better performance (lower AIC) than either DTW with the raw data or the linear method with either the function estimates or their derivatives.

In §2.4 we referred to multiple-term models which, in the linear case, minimize criterion (10). For logistic regression with $R = 2$ terms and no scalar covariates, such a model might

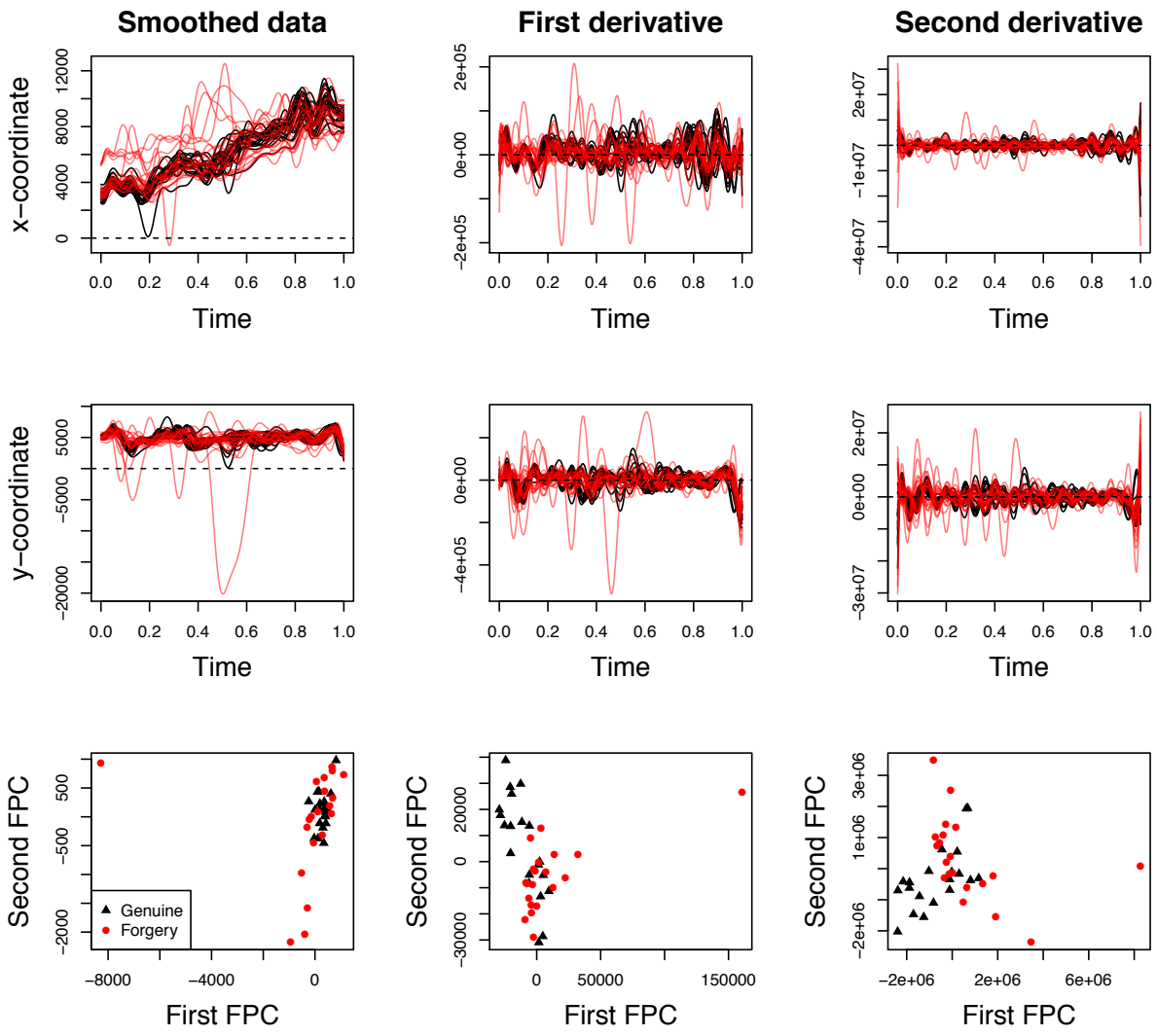

Figure 4: First row: Smoothed x-coordinate curves and their first and second derivatives (in the online version, the 20 true signatures appear in black and the 20 forgeries in red). Second row: Same, for the y-coordinates. Third row: Plots of first vs. second FPC scores.

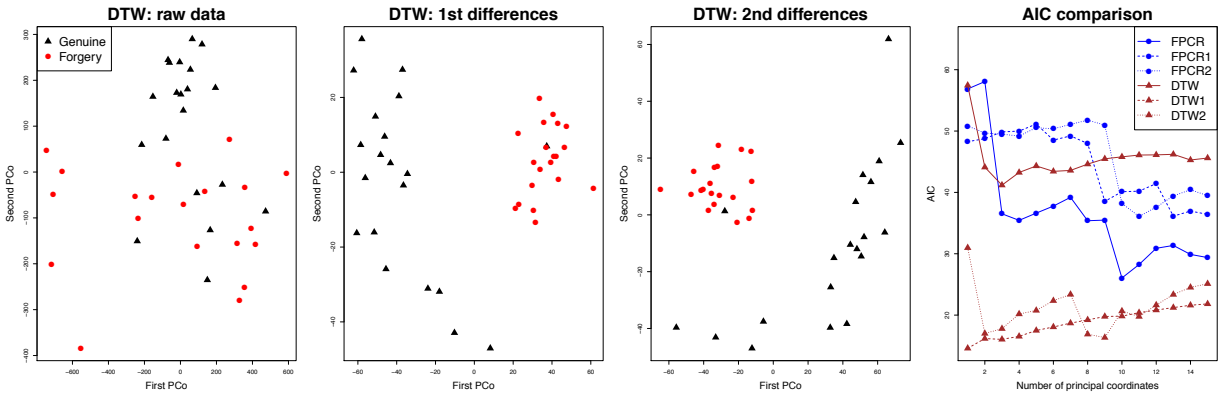

Figure 5: First three subfigures: First vs. second DTW-based principal coordinates for the original and differenced data. Last subfigure: AIC performance of logistic ridge FPCR (L^2-based PCoRR) vs. logistic DTW-based PCoRR for different numbers of PCo's. In the legend, "1" and "2" refer to first and second derivatives (for FPCR) or differences (for DTW).

be expressed as

$$\text{logit}(\boldsymbol{p}) = \boldsymbol{W}_{q_1}^{(1)}\boldsymbol{\gamma}_1 + \boldsymbol{W}_{q_2}^{(2)}\boldsymbol{\gamma}_2 \tag{20}$$

where $\boldsymbol{p} = [\Pr(y_1 = 1), \ldots, \Pr(y_n = 1)]^T$. To illustrate the potential utility of such models, in Figure 6 we display fitted values (estimated probabilities that the signatures are genuine) from logistic PCoRR models, with 5 PCo's, based on (i) L^2 distance among (i.e., FPCR with) the 2nd derivative curves, (ii) DTW distance among the 2nd-differenced data, and (iii) model (20) combining both of the above terms, i.e., $\boldsymbol{W}_{q_1}^{(1)} = \boldsymbol{W}_5^{(1)}$ based on L^2 distance and $\boldsymbol{W}_{q_2}^{(2)} = \boldsymbol{W}_5^{(2)}$ based on DTW distance. The first 20 and last 20 points, respectively, represent the genuine and forged signatures. Thus, according to the two-term model (iii), the estimated probability of being genuine is near 1 for all the true signatures and near 0 for all the forgeries. Such near-perfect in-sample prediction might, *a priori*, result from overfitting; but the AIC value for this model is lower than for either of the single-term models, i.e., the two-term model appears to be best in terms of out-of-sample prediction. We note that the AIC used here is the new modified criterion implemented in `mgcv`, which modifies the degrees of freedom to account for smoothing parameter uncertainty (Wood et al., 2015). The two-term model assigns more weight to the DTW-based than to the L^2-based PCo's (4.25 vs. 2.88 df), as one would expect given the superior performance of

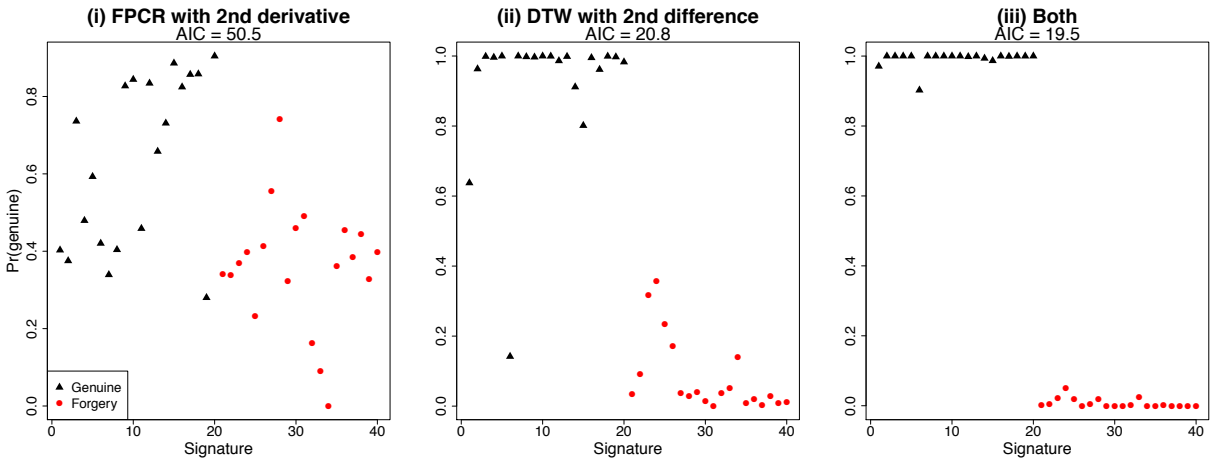

Figure 6: Fitted values from logistic PCoRR models based on (i) L^2 distance among 2nd-derivative curves (equivalent to logistic FPCR), (ii) DTW distance among 2nd-differenced curves, and (iii) model (20) incorporating both terms. In each case 5 PCo's are used.

the DTW-only vs. the L^2-only model.

5 Discussion

While our toy and real-data examples have involved DTW distance, PCoRR can equally well be applied with other distances among functional predictors, such as ISOMAP distance (Chen and Müller, 2012), the optimally weighted L^2 distance of Chen et al. (2014), or Mahalanobis distance (Galeano et al., 2015). PCoRR applies as well to non-functional data objects among which a distance or kernel is defined, as in the references cited in §3.3. Thus our contribution may be viewed either as a new approach to scalar-on-function regression, or more broadly as an extension of generalized additive models to incorporate data expressed as distances or kernels.

Our only requirements for the distance among data points are those given at the beginning of §2.1. A downside of this flexibility is that for non-Euclidean distances the matrix $\boldsymbol{K}$ in (4) is not positive semidefinite and thus mainstream kernel learning theory does not apply. It may, however, be possible to derive error results for PCoRR that combine kernel-based error bounds as in Steinwart and Christmann (2008) with bounds on the

approximation error due to truncation.

Another limitation of PCoRR is that it does not yield a coefficient function and hence is less interpretable than the functional linear model (1). It would therefore be helpful to have straightforward methodology for testing the null model (1) against the PCoRR alternative. We hope to present such methodology in a forthcoming paper.

Acknowledgments. The authors thank Lan Huo, Lei Huang, Huaihou Chen and Rong Jiao for their assistance in implementing the methods proposed here. Preliminary versions of this paper were presented at the Banff International Research Station workshop "Frontiers in Functional Data Analysis" in July 2015, and at the International Workshop on Advances in Functional Data Analysis held at Universidad Carlos III de Madrid in November 2015. We thank the participants of both workshops for their helpful feedback.

Appendix A: A note on interpolation formula (9)

At first glance, formula (9) has the appealing consequence that if we had $n^* = n$ new points such that $\boldsymbol{K}^* = \boldsymbol{K}$, the interpolated coordinates would be given by $\boldsymbol{W}_q^* = \boldsymbol{K}\boldsymbol{U}_q\boldsymbol{\Delta}_q^{1/2} = \boldsymbol{W}_q$. The equality $\boldsymbol{K}^* = \boldsymbol{K}$ does not, however, hold in general given n new observations that are identical to the original ones. What *is* true in that case, assuming $\delta_q > 0$, is that

$$\boldsymbol{K}^*\boldsymbol{U}_q = \boldsymbol{K}\boldsymbol{U}_q. \tag{21}$$

To see this, define $a_{rs} = -\frac{1}{2}d_{rs}^2$, $\bar{a}_r = \frac{1}{n}\sum_{s=1}^n a_{rs}$ and $\bar{a} = \frac{1}{n^2}\sum_{r=1}^n\sum_{s=1}^n a_{rs}$. Then $\boldsymbol{K} - \boldsymbol{K}^*$ has (m,i) entry $(a_{mi} - \bar{a}_m - \bar{a}_i + \bar{a}) - (a_{mi} - \bar{a}_i + \frac{1}{2}\bar{a}) = \frac{1}{2}\bar{a} - \bar{a}_m$. Thus the mth row of $\boldsymbol{K} - \boldsymbol{K}^*$ is $(\frac{1}{2}\bar{a} - \bar{a}_m)\mathbf{1}^T$, and that of $(\boldsymbol{K} - \boldsymbol{K}^*)\boldsymbol{U}_q$ is $(\frac{1}{2}\bar{a} - \bar{a}_m)\mathbf{1}^T\boldsymbol{U}_q = (\frac{1}{2}\bar{a} - \bar{a}_m)\mathbf{1}^T\boldsymbol{K}\boldsymbol{U}_q\boldsymbol{\Delta}_q^{-1} = \mathbf{0}^T$. Equation (21) follows, implying that $\boldsymbol{W}_q^* = \boldsymbol{W}_q$ and hence, as one would expect in this case, the predicted values $\hat{\boldsymbol{y}}^*$ are the same as the fitted values $\hat{\boldsymbol{y}}$.

Appendix B: Relating distances and kernels

Expression (4) can be viewed as establishing a rough correspondence between distance matrices $\boldsymbol{D}$ and kernel matrices $\boldsymbol{K}$. However, the kernel (or Gram) matrix encountered

in the kernel literature is generally positive semidefinite (indeed, it is generally given by $[k(x_i, x_j)]_{1 \le i,j \le n}$ for some well-defined function $k : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$), whereas $\boldsymbol{K}$ in (4) need not be. A fundamental result (Theorem 14.2.1 of Mardia et al., 1979) says that $\boldsymbol{K}$ is positive semidefinite if and only if $\boldsymbol{D}$ is Euclidean, i.e., there exist n points in a Euclidean space such that for each i, j, d_{ij} equals the Euclidean distance between the ith and jth points.

Working in the opposite direction, given a positive semidefinite kernel matrix $\boldsymbol{C}$, the standard transformation (13) defines a corresponding distance matrix $\boldsymbol{D}$. As noted in the proof of Theorem 1, the matrix $\boldsymbol{K}$ (4) associated with that distance matrix is equal to $\boldsymbol{HCH}$, i.e., to the centered version of the original kernel matrix $\boldsymbol{C}$ (Shawe-Taylor and Cristianini, 2004).

In summary, (4) and (13) establish a one-to-one correspondence between *Euclidean* distance matrices and *centered* kernel matrices.

It is worth noting that when $\boldsymbol{K}$ is defined as a kernel matrix—as opposed to being derived from a distance matrix via (4)—the matrix $\boldsymbol{W}_q$ defined as in §2.2 gives the leading *kernel principal components* of the data (Schölkopf et al., 1998). For further discussion of the connection between distances and kernels, see Schölkopf (2001) and Faraway (2012).

SUPPLEMENTARY MATERIAL

R code for analyses: Code to reproduce the analyses of the toy data and the signature verification data. (GNU zipped tar file)

References

Aguilera, A. M., F. A. Ocaña, and M. J. Valderrama (1999). Forecasting with unequally spaced data by a functional principal component approach. *Test 8*(1), 233–253.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pp. 267–281. Akademiai Kiado.

Baíllo, A. and A. Grané (2009). Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis 100*(1), 102–111.

Boj, E., A. Caballé, P. Delicado, A. Esteve, and J. Fortiana (2015). Global and local distance-based generalized linear models. *TEST*, in press.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika 48*(2), 305–308.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics and Probability Letters 45*(1), 11–22.

Chen, D. and H.-G. Müller (2012). Nonlinear manifold representations for functional data. *Annals of Statistics 40*(1), 1–29.

Chen, H., P. T. Reiss, and T. Tarpey (2014). Optimally weighted L^2 distance for functional data. *Biometrics 70*, 516–525.

Corrada Bravo, H., K. E. Lee, B. E. Klein, R. Klein, S. K. Iyengar, and G. Wahba (2009). Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences 106*(20), 8128–8133.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*(4), 317–403.

Cuadras, C. and C. Arenas (1990). A distance based regression model for prediction with mixed data. *Communications in Statistics—Theory and Methods 19*(6), 2261–2279.

Cuadras, C. M., C. Arenas, and J. Fortiana (1996). Some computational aspects of a distance-based model for prediction. *Communications in Statistics—Simulation and Computation 25*(3), 593–609.

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis 12*(1), 136–154.

Eckart, C. and G. Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika 1*(3), 211–218.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science 11*(2), 89–121.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Faraway, J. J. (2012). Backscoring in principal coordinates analysis. *Journal of Computational and Graphical Statistics 21*(2), 394–412.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer.

Galeano, P., E. Joseph, and R. E. Lillo (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics 57*(2), 281–291.

Geenens, G. (2011). Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys 5*, 30–43.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software 31*(7), 1–24.

Gönen, M. and E. Alpaydın (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research 12*, 2211–2268.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*(3-4), 325–338.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika 55*(3), 582.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Boca Raton, FL: Chapman & Hall.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. New York: Springer Science & Business Media.

Houmani, N., A. Mayoue, S. Garcia-Salicetti, B. Dorizzi, M. I. Khalil, M. N. Moustafa, H. Abbas, D. Muramatsu, B. Yanikoglu, A. Kholmatov, M. Martinez-Diaz, J. Fierrez, J. Ortega-Garcia, R. Alcobé, J. Fabregas, M. Faundez-Zanuy, J. M. Pascual-Gaspar, V. Cardeñoso Payo, and C. Vivaracho-Pascual (2012). BioSecure signature evaluation campaign (BSEC'2009): Evaluating online signature algorithms depending on the quality of signatures. *Pattern Recognition 45*(3), 993–1003.

Hyndman, R. J. and H. L. Shang (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics 19*(1), 29–45.

Kholmatov, A. and B. Yanikoglu (2005). Identity authentication using improved online signature verification method. *Pattern Recognition Letters 26*(15), 2400–2408.

Lu, F., S. Keleş, S. J. Wright, and G. Wahba (2005). Framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences of the United States of America 102*(35), 12332–12337.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, New York.

Marx, B. D. and P. H. C. Eilers (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics 41*(1), 1–13.

Miller, D. L. and S. N. Wood (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics 21*(4), 715–731.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *Annals of Statistics 33*(2), 774–805.

Müller, H.-G. and F. Yao (2008). Functional additive models. *Journal of the American Statistical Association 103*, 1534–1544.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications 9*(1), 141–142.

Preda, C. (2007). Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference 137*(3), 829–840.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.). New York: Springer.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2014). *fda: Functional Data Analysis*. R package version 2.4.4.

Randolph, T. W., J. Harezlak, and Z. Feng (2012). Structured penalties for functional linear models—partially empirical eigenvectors for regression. *Electronic Journal of Statistics 6*, 323.

Reiss, P. T., L. Huang, and M. Mennes (2010). Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics 6*(1), article 28.

Reiss, P. T. and R. T. Ogden (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association 102*, 984–996.

Reiss, P. T. and R. T. Ogden (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B 71*(2), 505–523.

Reiss, P. T. and R. T. Ogden (2010). Functional generalized linear models with images as predictors. *Biometrics 66*(1), 61–69.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Sakoe, H. and S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*(1), 43–49.

Schölkopf, B. (2001). The kernel trick for distances. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pp. 301–307. MIT Press.

Schölkopf, B., R. Herbrich, and A. J. Smola (2001). A generalized representer theorem. In D. Helmbold and B. Williamson (Eds.), *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001*, pp. 416–426. Berlin and Heidelberg: Springer.

Schölkopf, B., A. Smola, and K.-R. Müller (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation 10*(5), 1299–1319.

Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. New York: Cambridge University Press.

Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. New York: Springer Science & Business Media.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial Mathematics.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya A 26*, 359–372.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B 65*(1), 95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B 73*(1), 3–36.

Wood, S. N., N. Pya, and B. Säfken (2015). Smoothing parameter and model selection for general smooth models. arXiv preprint arXiv:1511.03864.